Match Predictions in the National Hockey League using Box Scores

Michael John Davis
Simon Fraser University

Tim Swartz
Simon Fraser University

Oliver Schulte
Simon Fraser University

Juan Camilo Gamboa Higuera
Sportlogiq

Mehrsan Javan
Sportlogiq

September 10, 2021

Abstract

Sports predictions are typically improved by incorporating as much relevant information as possible. In this paper we show a computationally straightforward method for including box score data that are available prior to a match. Box score data provide team summary statistics in a match (e.g. the number of shots on target, faceoffs won, etc.), and are a source of information about team performance. Our main idea is to treat event box scores like win/loss outcomes, and compute event ratings for each team using the Glicko system. For example, each team is assigned a shot rating, a face-off rating, etc. The event ratings are informative in themselves, and can be used as covariates to improve the accuracy in a match prediction model. We illustrate the approach based on matches in the National Hockey League.

1 Introduction

With the globalization of sport and the relaxation of restrictions

related to sports gambling (see www.actionnetwork.com/news/legal-sports-betting-united-states-projections), it is clear that interest in the prediction of sporting outcomes will only increase over time and across sports.

The sport for which sports gambling is the most popular is undoubtedly association football (i.e. soccer). Methods of prediction have been well researched in the literature, where a small sample of recent publications include Wheatcroft (2020), Constantinou (2019) and Groll et al. (2019).

In the sport of hockey, the National Hockey League (NHL) is the premier league in the world. Whereas gambling on match outcomes in the NHL is available on major internet betting platforms (e.g. Pinnacle, bet365, Bodog, Ladbrokes, etc), academic research on NHL prediction is not as extensive as in soccer. The methods that have been developed for prediction tend to fall within simpler regression frameworks; see for example, Buttrey (2016), Marek et al. (2014) and Stanek (2017). In this paper, we propose methods which provide pre-match probability estimates for NHL match prediction.

A key factor in the success of a prediction system is the incorporation of as much relevant information as possible. For our purposes it is helpful to distinguish between three levels of information arising from matches. To be specific, suppose we wish to make a prediction for the match between the Colorado Rockies and the Las Vegas Golden Knights on June 10, 2021.

(1) Final match outcomes only: For each team, match results are readily available from on-line sources. We know how many games Colorado and Las Vegas won prior to June 10, and against whom. Match outcome information is the basis of many seminal rating systems including Elo (1986), Thurstone (1927) and Bradley and Terry (1952).

(2) Summary covariates: There are many events in ice hockey which contribute to winning. For example, shots-on-goal, possession, penalty avoidance and body-checks all contribute to winning. However, summary covariates (taken over all games prior to the match of interest) do not take into account the idiosyncrasies involving specific team matchups and the result of the duals between pairs of teams. Intransitivity between teams in various sports has been investigated by Chen and Joachims (2016).

(3) Contextual covariates via box scores: For every match, an official NHL box score records the outcomes of events between two teams, and hence provides more detail than summary data. For example, a box score provides statistics such as the number of shots by a team and the percentage of face-offs they have won in a particular match. So we know the shot history of Colorado and Las Vegas in each of their games prior to June 10, and against whom.

Various data collecting organizations such as Sportlogiq provide enhanced box scores with information that goes beyond the official NHL box scores. Such box scores may be derived from play-by-play event data or potentially from tracking data. Box scores for the match of interest are unknown prior to the match, and therefore cannot be used for match prediction. In our example, we do not know in advance how many shots Colorado will manage against Las Vegas. The incorporation of past box scores in match prediction algorithms requires a novel approach that determines and utilizes events from previous matches. In this paper, we utilize the Glicko ratings (Glickman 1999) as box score covariates. While we focus on ice hockey as our target application, our method can be implemented using box score data from any sport.

The basic idea of our approach is that a separate Glicko rating is computed for each box score event. For example, our method computes a shot rating for Colorado, a shot rating for Las Vegas, a face-off rating for Colorado, a face-off rating for Las Vegas, etc. The ratings are computed from all of the box scores prior to the match on June 10. Therefore, the ratings provide a data reduction. That is, the ratings are a synthesis of box score results from the individual matches. And a key aspect of this is that box scores results are contextual in the sense that the opponents are known. Due to the Glicko construction, the difference between two team ratings can be interpreted as a prediction of their expected share of the box score event. For example, the difference in the shot rating between Colorado and Las Vegas predicts each teams percentage of the total shots in the match.

Context is an important aspect of our approach. A simple proportion of Colorados wins prior to June 10 neglects to take into account the opposition. Clearly, beating a strong team is more indicative of strength than beating a weak team. Similarly, a simple average of Colorados shots on target prior to June 10 neglects to take into account the teams against which Colorado managed the shots. Another aspect

of context is timing; recent matches are more indicative of strength than matches from the distant path. The Glicko system takes timing into account. The Glicko system is also dynamic as it captures the realism that team strength changes throughout a season.

Another feature of our approach is interpretability. Our event ratings represent the strength of a team for particular types of events relative to its opponents. For example, a teams face-off rating represents how good it is at winning face-offs. When a rating covariate is included in the prediction system (i.e. logistic regression), the corresponding fitted parameter allow us to interpret the extent to which the rating affects winning. The ratings are also of interest to coaches, fans, and other stakeholders, even outside our target application of match outcome prediction.

A final feature of our system is simplicity of computation and usage. We show how Glicko win ratings can be adapted for computing Glicko ratings for any box score event. The simplicity allows users to develop similar prediction systems for other sports whenever box scores are available.

We note that Glicko and its forebearer Elo are rating systems have been widely used in a variety of sports to predict match outcomes. To our knowledge, this is the first instance where Glicko is used in an intermediate step to derive dynamic covariates for the purpose of match prediction. In addition, we utilize play-by-play data in the estimation procedure; this level of match detail has also not been previously considered in NHL match prediction.

In Section 2, we describe our extensive data sources and the data management required. In Section 3, we explain how an event rating for each box score type can be computed using the Glicko algorithm. In Section 4, a logistic regression model is proposed where we provide details on how the event ratings are incorporated as covariates. We describe the process of variable selection which leads to our proposed model. The model is validated in Section 5 where a Brier score analysis is provided. Also, variations of the model are shown to be a profitable in the context of sports gambling. We conclude with a brief discussion in Section 6.

Event Description	Avg	Min	Max
goals scored	2.87	0	10
goals scored at even strength*	2.18	0	10
faceoff wins	29.03	11	53
shots on target*	31.46	12	65
shots wide of target*	29.62	8	61
checks or hits*	25.70	4	61
failed loose puck recoveries	91.60	30	204
successful breakouts*	11.38	0	38
successful offensive zone dump-ins*	32.09	10	60
unsuccessful offensive zone dump-ins	7.45	0	23
successful passes*	314.82	180	487
unsuccessful passes*	128.09	71	189
opponent shots blocked	14.19	1	39
opponent passes blocked*	62.14	19	112
penalties	3.19	0	10
penalties at even strength*	2.84	0	9

Table 1: A sample of event types from Sportlogiq play-by-play data. We list 16 event types from the original 40 available in the data. Summing each event type yields a box score event for each team and game. The table gives statistics taken over all box scores. Variables that are marked with asterisk are those that are included after variable selection in the final model (see Section 3).

2 Data

The Sportlogiq play-by-play data cover four seasons (2017-2018 through 2020-2021) and a total 4,877 games. They include 40 event types, including those shown in Table 1. The data are collected through a combination of manual video annotation and computer vision techniques. Based on domain knowledge, these events are believed to have an impact on winning matches. For a given team and game, we can up add up the event occurrences by the team to obtain a *game event count* that is one of the statistics listed in the box score for the game. Note that in some cases a large game event count is viewed as a positive contribution to winning. In other cases, a small game event count is viewed as a positive contribution to winning.

Some of these events are not independent but connected by their

semantics. For example, if a player passes the puck, this is recorded as both a successful pass and a successful reception.

3 Event-based Glicko Ratings

We now explain how an event rating for each box score type can be computed using the Glicko algorithm. There exist alternative versions of Glicko; we present a scaled down version corresponding to the Glicko 1.0 system (Glickman 2016) with simplifications and settings determined to suit our application. It is interesting to note that Elo can be described as a special case of Glicko. Our general framework for summarizing past box scores can be applied with any rating system.

In our framework, prior to a match, the team of interest has a Glicko rating r and a Glicko rating deviation RD for each box score event. These dynamic parameters represent team strength and variability, corresponding to the event. Denote the opponent's parameters by r_o and RD_o, respectively. At the beginning of data collection, teams are assigned starting values $r = 1500$ and $RD = 350$.

After the match, a box score outcome $p \in (0, 1)$ is obtained which characterizes the proportion of the event realized by the team of interest. The introduction of p is a twist on Glicko where Glicko outcomes in a paired comparisons setting are restricted to 1, 1/2 and 0 corresponding to a win, loss and draw, respectively. More sophisticated uses of margins of victory in Elo have been developed by Kovalchik (2020). With the box score outcome p observed for a particular event, the team's Glicko parameters are updated for the pair of teams as follows:

$$r' = r + \frac{\ln(10)/400}{1/RD^2 + 1/d^2}\, g(RD_o)(p - E(p)) \quad (1)$$

$$RD' = (1/RD^2 + 1/d^2)^{-1/2} \quad (2)$$

where

$$\begin{aligned} g(RD) &= \left(1+\frac{3(RD\ln(10))^2}{(400\pi)^2}\right)^{-1/2} \\ E(p) &= (1+10^{-g(RD_o)(r-r_o)/400})^{-1} \\ 1/d^2 &= (g(RD_o)\ln(10)/400)^2 E(p)(1-E(p)) \ . \end{aligned}$$

In Glicko 1.0, there is a feature where greater variability (i.e. larger RD) occurs when teams have played fewer matches and where matches are played in the distant past. This is important for our application since team compositions tend to change during the off-season due to player trades and the draft. Values of RD at the beginning of a season should be larger than during mid-season. Although we make use of the Glicko timing feature, we have not outlined the details in the description above.

Therefore, entering a match, each team will have Glicko ratings $r_1, r_2, \ldots, r_k$ and Glicko rating deviations $RD_1, RD_2, \ldots, RD_k$ corresponding to the k box score events. These ratings are updated after each box score result using equations (1) and (2).

4 Logistic Regression Match Prediction

In our research, we have considered various models that utilize Glicko ratings for events. In this section, we describe a basic logistic regression model for prediction, and we indicate directions where alternative models may be considered.

Let Y be the binary variable corresponding to whether the home team wins. $Y \sim \text{Bernoulli}(p)$ where the logit function describes the linear relationship between the probability $p = \text{Prob}(Y = 1)$ and covariate vectors. Specifically, we define

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_Z Z + \beta_h X_h + \beta_r X_r. \qquad (3)$$

In (3), the covariate Z is a vector of situational variables where the default case corresponds to $Z = 0$. For example, we define $Z_1 = 0$ if the home team is using their main goaltender, $Z_2 = 0$ if the road

team is using their main goaltender, Z_3 is the days of rest for the home team, and Z_4 is the days of rest for the road team.

The covariate vectors X_h and X_r in (3) correspond to the home and road teams, respectively, and are obtained from the Glicko ratings developed in Section 3. Specifically, for each event type with an asterisk in Table 1, we include the team's Glicko rating as a covariate. A modification that we have introduced is based on the consideration of team rosters. If a player is absent on game day, then the Glicko ratings are adjusted according to his individual contribution to the corresponding events.

The total number of parameters in the model (3) is $2|X| + |Z| + 1$ where $|X|$ is the number of event types included by variable selection and $|Z|$ is the number of situational covariates. For the Sportlogiq dataset, $|Z| = 4$ and $|X| = 10$, so the total number of parameters is 25. Note that the selected parameters are the same for each team. The model may be extended to account for team-specific effects. Introducing two parameters for each event rating captures interactions between the home/road status of a team and event ratings. For example, a β_h parameter for the face-off rating of the home team and a separate β_r parameter for the face-off rating of the road team can represent that winning face-offs at home may be more important than winning them on the road. In simulations we found that capturing interactions between home/road status and event ratings substantially improved the predictive accuracy of the model, compared to using the difference in event ratings with a single parameter β.

The model is easily fit using the *bayesglm* R package with default settings. Compared to the *glm* package, the Bayesian version incorporates a prior, which we found improves numerical stability and slightly more accurate predictions. When the model is fit and estimates are obtained, the probability that $Y = 1$ for a particular covariate pattern is given by

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_Z Z + \hat{\beta}_h X_h + \hat{\beta}_r X_r)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_Z Z + \hat{\beta}_h X_h + \hat{\beta}_r X_r)} \ .$$

From the long list of potential Glicko covariates described in Table 1, only those 10 variables with asterisks were included in the final model. Variable selection was determined by fitting the model using

training data from the two seasons $2017-2018$, $2018-2019$ and the first 20% of matches from $2019-2020$. The remainder of the $2019-2020$ served as test data. We selected the subset of event types through a search that led to greatest predictive accuracy on the test data.

5 Model Evaluation

One of the most common bets in hockey is the *moneyline*. In a moneyline wager, you are simply betting on which team will win, and in hockey, there is always a winner at the end of the game.

5.1 Predictive Accuracy

Models were fit using three seasons worth of data. The 2020-2021 season was held out to assess predictive accuracy. We compare two prediction methods.

(1) Box Score Ratings: We predict a home win ($Y = 1$) using the logistic regression model (3) which includes the box score ratings X.

(2) Win-based Glicko: We predict a home win using logistic regression which includes the Glicko 1.0 rating covariate (Glickman 2016) based on win-loss outcomes and the the situational covariates Z. The box score event covariates X are excluded.

5.1.1 Training Set Results

Table 2 shows the model fit for the training seasons. We report accuracy (the percentage of correctly predicted winners) and the Brier score (Brier 1950) which is a squared-error metric for which lower values are better. Utilizing box score ratings allows the model to fit the training data substantially better than win-loss outcomes only. Both methods fit the 2019-2020 season worse than previous seasons; this may be related to the COVID-19 suspension that occurred in March 2020.

5.1.2 Test Set Results

We computed test predictions on season 2020-2021 in an incremental setting (Carpenter 2017). When making predictions for a game on day $m + 1$, we fit parameters $\hat{\beta}_m$ based on the games played on days

Box Score Ratings			Win-based Glicko	
Season	Accuracy	Brier Score	Accuracy	Brier Score
2017-2018	0.559	0.244	0.532	0.249
2018-2019	0.579	0.244	0.527	0.250
2019-2020	0.526	0.250	0.511	0.250

Table 2: *Training Set* comparison of Box Score Ratings with Win-based Glicko baseline where the results are rounded to 3 digits.

Box Score Ratings			Win-based Glicko	
Season	Accuracy	Brier Score	Accuracy	Brier Score
2020-2021	0.583	0.244	0.560	0.247

Table 3: *Test Set* comparison of Box Score Ratings with Win-based Glicko baseline where the results are rounded to 3 digits.

$\leq m$. Thus ratings and parameters are both updated dynamically.

An incremental approach takes into account that match outcomes are more like a time series than i.i.d. data, because match outcomes for different days are dependent. It is also more realistic as a betting scenario, where a gambler would at any point use all the information available to make the best bet. For example, if the weights estimated from the first three seasons are not optimal for season four, a rational gambler would adjust the weights during the season rather than keep using the same weights and lose money. While computationally efficient incremental methods are available for dynamically adjusting logistic regression weights in an on-line setting (Carpenter 2017), we simply applied the *bayesglm* package in batch mode, and input on all previous match data for dates $m = 1, \ldots$.

Table 3 shows the predictive accuracy for incremental learning with Box Score Ratings versus Win-based Glicko. The accuracy of both methods increases in the final season, which is evidence that each is using the data to improve predictions. The baseline Glicko method achieves its best Brier score in the most recent season, whereas the box score based method has a stable Brier score around 2.44 (except for the anomalous 2019-2020 season). The biggest advantage of the box score method is in beating the market, which we discuss next.

	Box Score Ratings				Win-based Glicko	
threshold ε	#bets	total profit	ROI	#bets	total profit	ROI
0.0	839	\$ 35.90	4.28%	839	-\$17.98	-2.14%
0.05	550	\$ 26.52	4.82%	520	-\$30.92	-5.95%
0.1	306	\$ 20.65	6.75%	267	-\$25.00	-9.36%

Table 4: Betting results on the most recent season 2020-2021. We report results for a simulated betting strategy that wagers \$1 whenever the difference between the model probabilities and the probabilities implied by the market odds exceeds a threshold. ROI is the profit divided by the number of bets.

5.2 Gambling Application

To assess the betting impact of Box Score Ratings, we examine a simple constant stake betting strategy with a threshold ε.

1. Let o be the bookmaker's implied odds and p be the model probability for both the home and road teams
2. If $|p - o| > \varepsilon$, wager \$1 on the team (home/road) for which the bookmaker's implied probability is underestimated

It is important to keep in mind that bookmaker's odds include a vigorish that makes it difficult to develop a profitable system. Bookmaker's odds are widely thought to be efficient (Gandar, Zuber and Johnson 2004); were they not so, bookmakers would struggle to stay in business.

Table 4 shows the total money won or lost by each method when used with a constant stake betting strategy during the 2020-2021 season, depending on three thresholds. Given that our model is efficient, the intuition is that higher thresholds identify cases where the bookmaker's odds are less reliable. The results were computed using the closing (i.e. final) odds from the Betfair website. These results provide strong evidence for the ability of the Box Score Ratings method to beat the market. More dramatic results may be realized with departures from fixed stake betting. For example, Chu, Wu and Swartz (2019) examine various strategies related to the Kelly criterion.

- There is a big difference in total profit. Betting \$1 on every game, the box score method gains \$35.90, whereas the baseline

method loses almost half that amount.

- The box score probabilities deviate from the market probabilities more strongly than with Win-based Glicko. For example, for the box score method there are 306 matches where the difference is at least 0.1, but for the baseline method, there are only 267 matches.

- The greater the difference between box score and market probabilities, the greater the return on investment (e.g., 6.75% for the 0.1 threshold). This is evidence that when the box score model deviates from the market, it tends to deviate in the right direction. In contrast, betting on matches where the baseline Glicko method deviates from the market exacerbates the losses. For the Box Score Ratings method, there is a tradeoff between a profitable ROI with many available matches and an even more profitable ROI with fewer matches for betting.

6 Conclusions

We have described a match prediction system for the NHL which leverages box scores to improve match outcome prediction. The key idea is to introduce Glicko ratings, one for each selected statistic in the box score, as covariates in a logistic regression model. We have demonstrated on NHL data that the system provides more accurate predictions than a traditional Glicko prediction system based on win-loss outcomes (and no box score data). The biggest performance difference is observed when the box score method is utilized as a gambling system: Simulating betting on the most recent NHL season, we observed that the model probabilities differ markedly from market odds, and the more they differ, the greater the return on investment from (constant-stake) bets. The simplicity of the system is appealing and permits development in other sports that provide box scores.

There are many ways that the system may be modified and possibly improved in future research. Some of these directions include:

- combine win-loss ratings with box score ratings

- leverage box score information to predict more complex outcomes (e.g. regular time versus overtime match scores, goal dif-

ferentials)

- include team specific effects
- identify new box score events that are predictive
- use modern machine learning algorithms as alternatives to logistic regression

In sum, our work has shown that team ratings based on box scores are a useful and computationally straightforward addition to the match prediction toolkit.

7 Bibliography

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39, 324-345.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1-3.

Buttrey, S.E. (2016). Beating the betting market on NHL hockey games. *Journal of Quantitative Analysis in Sports*, 12, 87-98.

Carpenter, B. (2016). Lazy sparse stochastic gradient descent for regularized multinomial logistic regression. Technical Report *Alias-i, Inc.* 2017.

Chen, S. and Joachims, T. (2016). Modeling intransitivity in matchup and comparison data. *Proceedings of the ACM International Conference on Web Search and Data Mining*, 227-236.

Chu, D., Wu, L. and Swartz, T.B. (2018). Modified Kelly criteria. *Journal of Quantitative Analysis in Sports*, 14, 1-11.

Constantinou, A.C. (2019). Dolores: A model that predicts football match outcomes from all over the world. *Machine Learning*, 108, 49-75.

Elo, A.E. (1986). *The Rating of Chess Players, Past and Present, Second Edition*, Arco Publishing, New York.

Gandar, J.M., Zuber, R.A. and Johnson, R.S. (2004). A reexamination of the efficiency of the betting market on National Hockey League

games. *Journal of Sports Economics*, 5, 152-168.

Glickman, M.E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, 48, 377-394.

Glickman, M.E. (2016). The Glicko system. Retrieved on August 5, 2021 from http://www.glicko.net/glicko.html

Groll, A., Ley, C., Schauberger, G. and van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15, 271-287.

Hoffmann, M.D., Loughead, T., Dixon, J.C. and Crozier, A.J. (2017). Examining the home advantage in the National Hockey League: Comparisons among regulation, overtime and the shootout. *Psychology of Sport and Exercise*, 28, 24-30.

Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. *International Journal of Forecasting*, 36, 1329-1341.

Marek, P., Sediva, B. and Toupal, T. (2014). Modeling and prediction of ice hockey match results. *Journal of Quantitative Analysis in Sports*, 10, 357-365.

Stanek, R. (2017). Home bias in sport betting: Evidence from Czech betting market. *Judgment and Decision Making*, 12, 168-172.

Swartz, T.B. and Arce, A. (2014). New insights involving the home team advantage. *International Journal of Sports Science and Coaching*, 9, 681-692.

Thurstone, L.L. (1927). A law of comparative judgment. The method of paired comparisons for social values. *Psychological Review*, 34, 368-389.

Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 36(3), 916-932.